HiTC - Exploration of High Throughput 'C' experiments

N. Servant, B. Lajoie, E. Nora, L. Giorgetti, C. Chen, J. Dekker, E. Heard, E. Barillot

August 10, 2012

Supplementary Data

This document present the main analysis steps covered by the HiTC package, and the complete R code used to perform them. The purpose of this document is not to explain in details the use of the HiTC package. For this, please read the vignette supplied with the package.

1 High-Throughput 'C' analysis workflow

While the use of high-throuput 'C' techniques is expected to increase in the coming years, it also creates some new statistical and bioinformatics challenges.
The Figure S1 summarizes the different steps of a classical 5C or Hi-C analysis.

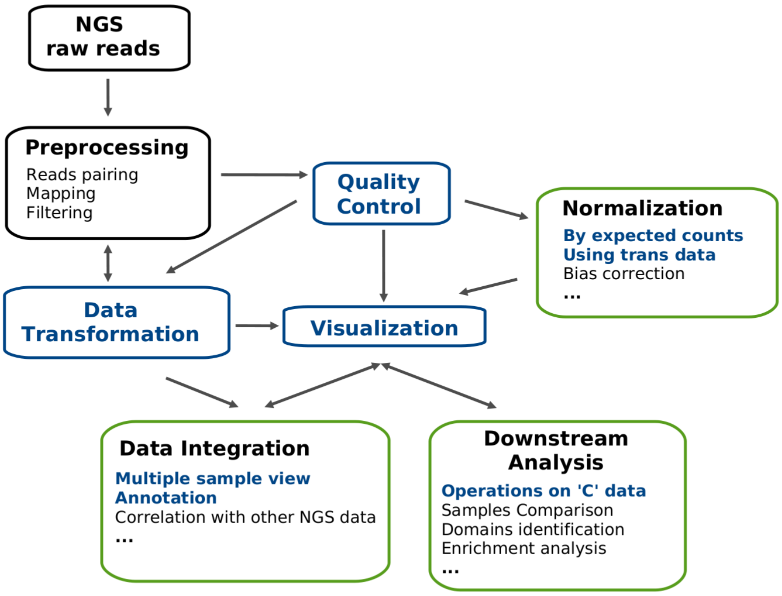

Figure S1: *Summary of the different steps of a classical high-throughput 'C' analysis. The steps in blue are covered by the current version of the HiTC package (version 1.1.2).*

The data visualization provided by the HiTC package is a central part of the analysis, as well as the data transformation, and the data normalization. The dowstream analysis and

data integration are then more difficult to generalize and are often closely related to the biological context of the project.
The parts in blue are the different steps of the analysis covered by the current version of the package. The HiTC package offers functionnalities in all the main analysis steps. In green are the futur fields of developments for this type of analysis for which the HiTC package already provides some functions.
While 5C enables analysis of interactions between many loci, it also required an extensive number of primers, which is not suitable for a genome-wide analysis as the Hi-C. Thus, the pre-processing of these two types of data is totally different with, for instance, two different mapping strategies and is not currently part of the package.

2 Load 5C dataset (Nora et al. 2012)

Nora et al recently used chromosome conformation capture carbon-copy (5C) in order to analyse the spatial organization of a 4.5 megabases region of the X inactivation center. Two Mouse samples, male undifferentiated ES cells (E14, GSM873935) and male embryonic fibroblasts (MEF, GSM873924) are included in the HiTC package.

```
> library(HiTC)
```

```
> data(Nora_5C)
> ## List of HTCexp objects describing the 5C dataset
> show(E14)
```

```
$chrXchrX
HTC object
Focus on genomic region [chrX:98831149-103425150]
CIS Interaction Map
Matrix of Interaction data: [511-528]
511  genome intervals from  chrX  ('ygi' object)
528  genome intervals from  chrX  ('xgi' object)
```

```
> show(MEF)
```

```
$chrXchrX
HTC object
Focus on genomic region [chrX:98831149-103425150]
CIS Interaction Map
Matrix of Interaction data: [511-528]
511  genome intervals from  chrX  ('ygi' object)
528  genome intervals from  chrX  ('xgi' object)
```

3 Quality Control

The first step after data pre-procesing is a quality control to check whether the data are likely to reflect cis and/or trans chromosomal interactions rather than just random collisions.

The HiTC package allows to assess the percentage of reads aligned to interchromosomal and intrachromosomal interactions, and the distribution of the interaction frequency against the genomic distance between two loci (see Figure S2).

```
> CQC(E14)
```

```
       nbreads nbinteraction averagefreq medfreq
all   21241622        196642     108.022      20
cis   21241622        196642     108.022      20
trans        0             0       0.000       0
```

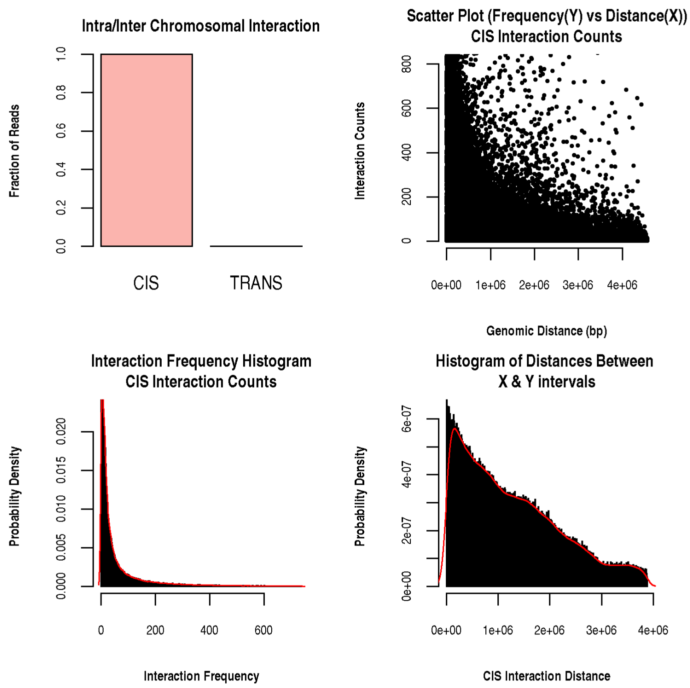

Figure S2: *Quality Control of 5C data. Top-left : proportion of inter/intra chromosomal interactions. Top-right : scatter-plot of interaction counts versus genomic distance between two loci. Bottom-rigth : histogram of interaction counts. Bottom-left : histogram of distances between two loci.*

4 Visualization of Interaction Maps

The interaction map represents the frequency at which each pair of restriction fragments have been ligated together during the 3C procedure. The goal is to visualize at once these counts for many pairs of restriction fragments across a large genomic region. Each entry in the matrix corresponds to a count information, i.e., number of times two restriction fragments have been sequenced as a pair.
Therefore 5C and Hi-C interaction maps are typically displayed using two dimensional heatmaps (see Figure S3) or a triangle view. The latest is particulary useful for interaction maps comparison and alignment with genomic or epigenomic features (see Figure S7).

```
> mapC(E14$chrXchrX)
```

```
[1] "minrange= 1  - maxrange= 741"
```

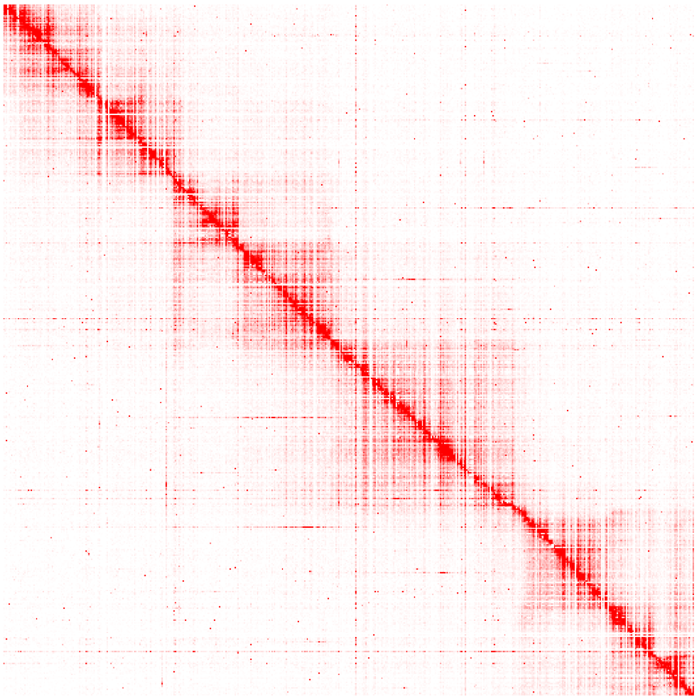

Figure S3: *5C interaction map of chromosome X.*

5 Data Transformation

5.1 Windowing

Each pixel of an interaction map can correspond either to a single restriction fragment, several restriction fragments or genomic intervals of any given size (and therefore various restriction fragment numbers).
To produce an interaction map, the genomic range of the display should be divided into appropriately size loci. This size depends on the resolution desired for the analysis. For instance, 5C data can be visualized at the primers resolution, or segmented into 100Kb or 1Mb bins that can be partially overlap or not. The binned cis interaction map is symmetrical around the diagonal (see Figure S4).

```
> ## Binning of 5C interaction map
> E14.binned <- binningC(E14$chrXchrX, binsize=100000, step=3)
> mapC(E14.binned)
```

```
[1] "minrange= 1  - maxrange= 385"
```

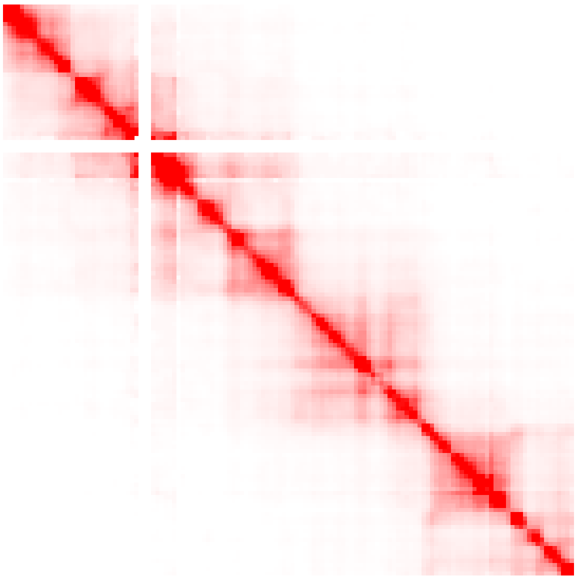

Figure S4: *Binned 5C interaction map of chrX:100295000-102250000.*

5.2 Data Normalization

Due to the polymer nature of chromatin, at small genomic distances, pairs of restriction fragments that are close to each other in the linear genome will give higher signal than fragments that are further apart. Such property leads to strongest counts falling on the heatmap diagonal. When considering any given pair of restriction fragments, it is therefore informative to assess whether the observed counts are above what is expected given the genomic distance that separate them.
Different ways of normalization have been proposed. In the current version of the package we propose to estimate the expected interaction counts as presented in Bau et al. (2011). The expected value is the interaction frequency between two loci that one would expect based on a sole dependency on the genomic proximity of these fragments in the linear genome. This can be estimated using a Lowess regression model (see Figure S5).

```
> ## Look at exptected counts
> E14exp <- getExpectedCounts(E14$chrXchrX, plot=TRUE)
```

Thus, interaction frequencies can be then normalized for distance by dividing the observed value by the expected value. For the following example, we decided to focus on a subset of the original dataset (see Figure S6).

```
> ## Focus on a subset chrX:100295000:102250000
> E14sub<-extractRegion(E14$chrXchrX, chr="chrX", from=100295000, to=102250000)
> E14sub.binned <- binningC(E14sub, binsize=50000, step=3)
> mapC(E14sub.binned, maxrange=100)
```

```
[1] "minrange= 0.5  - maxrange= 100"
```

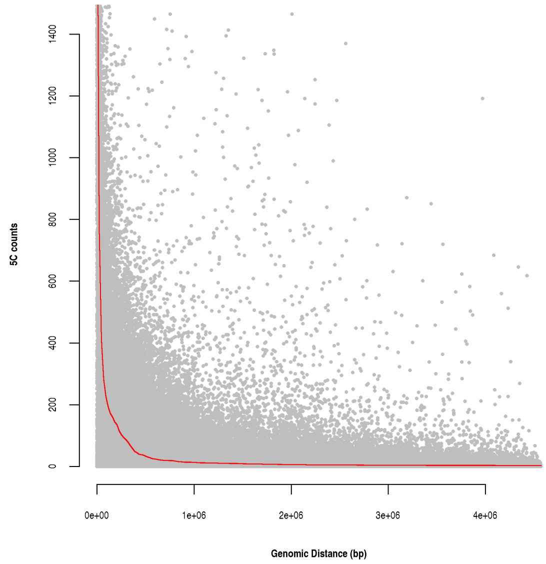

Figure S5: *Estimation of expected count using a Lowess smoothing.*

```
> ## Data normalization
> E14sub.norm <- normPerExpected(E14sub)
> E14sub.norm.binned <- binningC(E14sub.norm, binsize=50000, step=3)
> mapC(E14sub.norm.binned)
```

```
[1] "minrange= 0.133671  - maxrange= 2.978541"
```

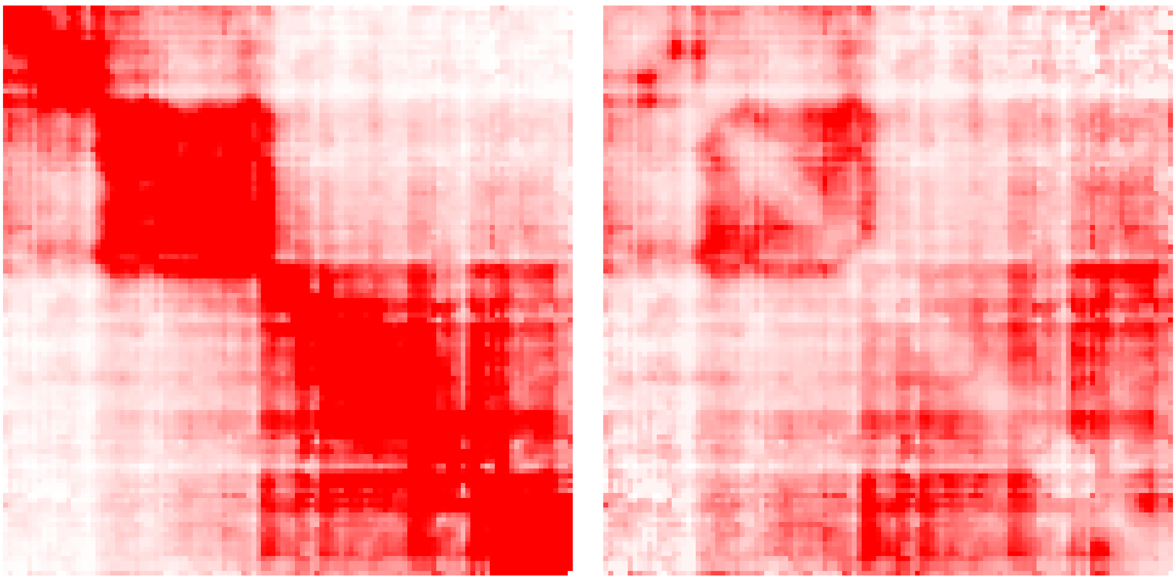

Figure S6: *Interaction maps of raw and normalized data.*

6 Annotation of Interaction Maps

The *HiTC* package contains functions for visualizing both genomic annotations (BED format) and the interaction maps (see Figure S7). For instance, the following example displays the CTCF enriched regions (Kagey et al. (2010)) and RefSeq genes over the interaction map of the E14 sample.

```
> E14.binned <- binningC(E14$chrXchrX, binsize=100000, step=3)
> exDir <- system.file("extdata", package="HiTC")
> Refgene <- readBED(file.path(exDir,"refseq_mm9_chrX_98831149_103425150.bed"))
> CTCF <- readBED(file.path(exDir,"CTCF_chrX_98892125_102969775.bed"))
> mapC(E14.binned,
+      giblocs=list(RefSeqGene=Refgene$Refseq_Gene, CTCF=CTCF$CTCF),
+      view=2)
```

```
[1] "minrange= 1  - maxrange= 385"
```

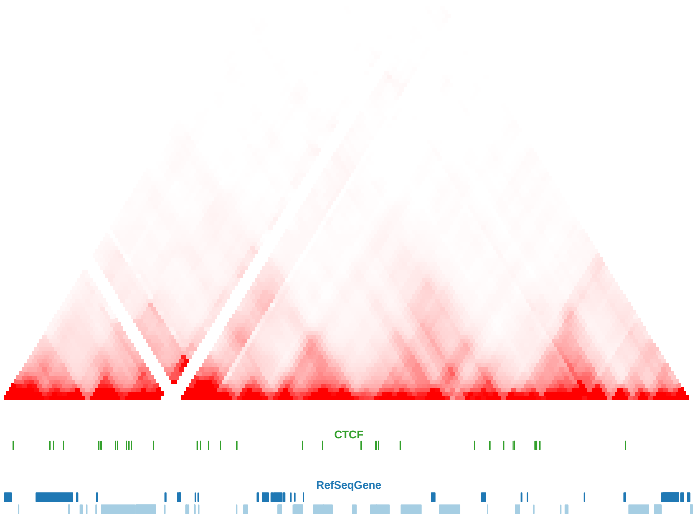

Figure S7: *Visualization of interaction map and genomic annotations.*

7 Comparison of E14 and MEF samples

The *HiTC* package provides methods to perform simple operations on *HTCexp*, such as dividing (see Figure S8), substracting two objects or extracting a genomic region.

```
> ## MEF sample normalization and binning
> MEFsub<-extractRegion(MEF$chrXchrX, chr="chrX", from=100295000, to=102250000)
```

```
> MEFsub.norm <- normPerExpected(MEFsub)
> MEFsub.norm.binned <- binningC(MEFsub.norm, binsize=50000, step=3)
> mapC(MEFsub.norm.binned)
```

```
[1] "minrange= 0.248574  - maxrange= 2.784829"
```

```
> ## log ratio MEF vs E14 interaction map
> mapC(divide(E14sub.norm.binned, MEFsub.norm.binned), log=TRUE)
```

```
[1] "minrange= 0.017179  - maxrange= 1.979405"
```

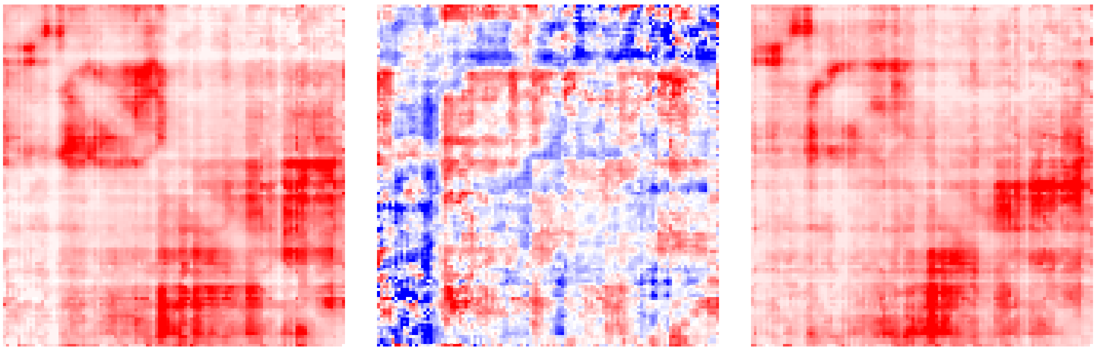

Figure S8: *From left to right; E14 normalized interaction map, Log ratio of normalized MEF and E14 interaction maps, and MEF normalized interaction maps.*

It also proposes a graphical view to compare two 'C' experiments. In the following example, the MEF sample is compared to the E14 sample (see Figure S9).

```
> ## Annotation and visualization of both samples
> MEF.binned <- binningC(MEF$chrXchrX, binsize=100000, step=3)
> mapC(E14.binned, MEF.binned,
+       giblocs=list(RefSeqGene=Refgene$Refseq_Gene, CTCF=CTCF$CTCF),
+       maxrange=100)
```

```
[1] "minrange= 0.5  - maxrange= 100"
[1] "minrange= 0.5  - maxrange= 100"
```

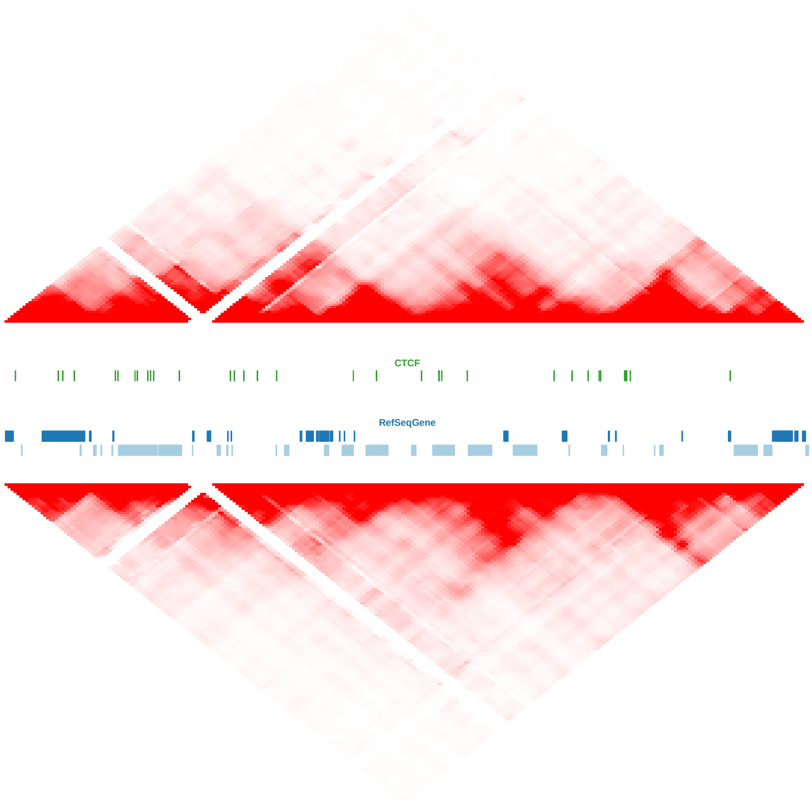

Figure S9: *Comparison of two binned interaction maps, and visualization with genomic annotations.*

Package versions

This document was generated using the following package versions:

- R Under development (unstable) (2012-06-24 r59622), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: BiocGenerics 0.3.0, Biostrings 2.25.6, genomeIntervals 1.13.2, GenomicRanges 1.9.31, girafe 1.9.0, HiTC 1.1.2, intervals 0.13.3, IRanges 1.15.20, lattice 0.20-6, latticeExtra 0.6-19, RColorBrewer 1.0-5, Rsamtools 1.9.21, ShortRead 1.15.9
- Loaded via a namespace (and not attached): Biobase 2.17.6, bitops 1.0-4.1, BSgenome 1.25.3, hwriter 1.3, stats4 2.16.0, tools 2.16.0, zlibbioc 1.3.0

References

D. Bau, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom. The three-dimensional folding of the Îś-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, 18(1):107–114, Jan 2011. doi: 10.1038/nsmb.1936. URL http://dx.doi.org/10.1038/nsmb.1936. 5

M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, Sep 2010. doi: 10.1038/nature09380. URL http://dx.doi.org/10.1038/nature09380. 7